Social desirability bias in candidate conjoint experiments

What is the optimal design when studying sensitive topics?

Dahl, Malte

Publication date:
2018

Social desirability bias in conjoint experiments: What is the optimal design when studying sensitive topics?

Malte Dahl *

Working paper

Abstract

An often-mentioned advantage of conjoint experiments over traditional survey experimental designs is that the former have the potential to mitigate social desirability bias. To what extent this is true may depend on a number of design choices – a concern that has received surprisingly little empirical attention. I conducted two studies in which I randomly assigned respondents to three types of conjoint designs in order to manipulate their awareness to sensitive features and possibilities for justifying inappropriate answers (N = 7,059). The results show that design variations significantly affect respondents' inferences about the research objective. However, there are no detectable differences between respondents' preferences across designs. This indicates that researchers using conjoint experiments should not compromise their choice of design to avoid social desirability bias.

*I would like to thank Mogens Jin Pedersen, Alex Coppock, Thad Dunning, Peter T. Dinesen, Frederik Hjorth, Kasper Moeller Hansen, Anders Woller, Ma Yi, Benjamin Egerod, Jens van der Ploeck and participants at the seminar *New approaches for the study of political behavior* at the Department of Political Science, University of Copenhagen. Also, a big thanks to Thomas Leeper for creating the R package *cregg*. Malte Dahl is PhD. candidate at the Department of Political Science, University of Copenhagen, 1353 Copenhagen (mrd@ifs.ku.dk).

Introduction

Conjoint experiments have become a standard part of the political science toolkit. These experiments are effective and low-cost tools that enable researchers to elucidate respondents' multidimensional preferences and test several causal hypotheses simultaneously (Hainmueller, Hopkins, and Yamamoto 2014; Hainmueller, Hangartner, and Yamamoto 2015; Bansak, Hainmueller, Hopkins, et al. 2017). Moreover, researchers can easily increase the effective sample size by letting each individual respondent answer several conjoints.[1]

Another considerable advantage that is often emphasised by proponents of conjoint experiments is that these designs have the potential to mitigate social desirability bias (SDB) (Hainmueller, Hopkins, and Yamamoto 2014; Horiuchi, Smith, and Yamamoto 2017; Teele, Kalla, and Rosenbluth 2018). The ability to obtain reliable answers is a key inferential issue in the survey-experimental literature and considering that conjoint experiments are often used to gauge respondents' reactions to sensitive dimensions, this is an essential quality. However, despite the prominence of conjoint designs, there has been surprisingly little effort to examine the extent to which, and the conditions under which, SDB is of concern when examining sensitive topics.

The perceived ability of conjoint experimental designs to mitigate SDB is grounded in two notions. First, since respondents are presented with numerous features, a given sensitive feature is 'masked' among other features that are also randomly varied (*attention assumption*). Therefore, it is argued, respondents cannot infer that the sensitive feature is of particular importance (Teele, Kalla, and Rosenbluth 2018). Second, respondents can always find multiple justifications for any given choice (Hainmueller, Hopkins, and Yamamoto 2014). This implies that inappropriate answers can be justified by (combinations of) the levels of other features in the experiment (*justification assumption*).

The extent to which these two assumptions hold may be heavily conditioned by a number of specific design choices. For example, there is a fundamental difference between paired and

[1] These designs have been used to study how voter preferences are shaped by political candidates' gender (Teele, Kalla, and Rosenbluth 2018) and class (Carnes and Lupu 2016), the way information on party affiliation moderates voter preferences (Kirkland and Coppock 2017), and Americans' attitudes towards immigrants (Hainmueller and Hopkins 2015).

single-profile designs, between designs that measure outcomes as a discrete choice, a rating or a combination of the two, and between designs that manipulate few or many features and feature levels.[2] Moreover, different randomisation schemes can be applied, with some studies randomising the number of features that are presented, randomising all or only some of the features and/or randomising feature levels with different probability weights (Hainmueller and Hopkins 2015). I argue, that these design differences are likely to have an effect on social desirability pressures because they influence (*i*) respondents' anticipation of the primary research objective, and (*ii*) the degree to which respondents can justify inappropriate answers over repeated tasks. I also argue that conjoint designs that, at least in theory, downplay social desirability pressures often compromise other important features of the experiment (e.g. statistical power or ecological validity). This raises an important question: what is the optimal design when studying sensitive topics in conjoint experiments?

In this pre-registered study[3], I aim to answer that question by randomly assigning respondents to seemingly similar conjoint designs that vary social desirability pressures. To do so, I ran two studies inspired by Sen (2017) and Hainmueller and Hopkins (2015), respectively. Both of these were conducted using Amazon's Mechanical Turk marketplace (N = 7,059). In each study, respondents were randomly assigned to one of three conjoint designs intended to either minimise or amplify their attention to sensitive dimensions and their possibilities for justifying inappropriate answers.[4] Specifically, in the first condition, *the high-contrast paired design*, each respondent was presented with a number of conjoint pairs in which the levels of a sensitive feature were repeatedly contrasted (e.g., a black vs. a white candidate). The second condition, *the restricted paired design*, was similar, except that the sensitive feature was only contrasted in a limited number of conjoint pairs. Finally, the third condition was a fully randomised *single-profile design* showing only one candidate at a

[2] *Features* can include, for example, age, party affiliation and gender, whereas *feature levels* are the values each feature can take, e.g., male/female in the case of gender.

[3] The project was registered at Open Science Framework and a pre-analysis plan of Study 1 can be found at www.osf.io/sf6h9, while a pre-analysis plan for study 2 can be found at www.osf.io/ket62

[4] This work is related to recent studies that have examined demand effects in survey experiments by inducing different degrees of information about the purpose of the study (Mummolo and Peterson 2018; De Quidt, Haushofer, and Roth 2017). However, instead of raising awareness of the research objective by providing respondents with explicit information, the present project sought to manipulate awareness to sensitive dimensions *through design*.

time.

The results demonstrate that these design differences significantly affect respondents' inferences about the research objective (i.e. their attention to a sensitive feature). Specifically, respondents assigned to a high-contrast paired conjoint design are much more likely to infer that the sensitive feature is the main focus of the study than respondents assigned to either of the other two designs. Surprisingly, and most importantly, the design differences do not translate into any immediate effect on respondents' priorities. When comparing the effects of the sensitive features across designs, there are no distinguishable differences: respondents' answers are stable. This evidence indicates that when researchers use conjoint designs to study sensitive topics, they should not compromise their choice of design due to the fear of SDB.

Social desirability bias in survey research

A common understanding of SDB is the respondent's lack of comfort to reveal his or her true attitudes (Tourangeau and Yan 2007; Groves et al. 2011; Holtgraves 2004). Respondents moderate their behavior by giving normatively positive responses in order to make themselves look more favourably and avoid the embarrassment, unease and distress that revealing socially undesirable answers may bring (Kaminska and Foulsham 2013). For example, respondents tend to underreport favoritism for preferred groups relative to nonpreferred ones (Janus 2010; Kuklinski et al. 1997) which leads to a misrepresentation of preferences.

Evidence on SDB in survey research generally suggests that it is a valid concern. This is demonstrated in studies that word questions in more or less threatening ways (Kuklinski et al. 1997), that change the interview setting (Krysan and Couper 2003), that compare results from list experiments with direct questions (Janus 2010; Gilens, Sniderman, and Kuklinski 1998) or studies that compare survey answers with register data (Hariri and Lassen 2017).

Moreover, several studies have demonstrated that some groups of respondents are more likely to provide socially desirable answers. For example, Berinsky and Lavine (2012) demonstrate that high self-monitors are more likely to offer socially acceptable answers. Other studies indicate that liberal

respondents are more likely to give untruthful answers to questions regarding race (Gilens, Sniderman, and Kuklinski 1998) and immigration restrictionist policy questions (Janus 2010). A related concern is that survey experiments frequently rely on online subject pools, like Amazon's Mechanical Turk, where experienced experimental participants have incentives to be especially attentive to researcher expectations (Krupnikov and Levine 2014). For this reason, Berinsky, Huber, and Lenz (2012) recommend that researchers avoid revealing their intentions in online survey experiments.[5]

Conjoint experiments as a means to overcome SDB

While SDB is a potential validity issue in all survey research, it is often claimed that conjoint experiments can mitigate some of these concerns (Hainmueller, Hopkins, and Yamamoto 2014; Liu 2018; Teele, Kalla, and Rosenbluth 2018). Two arguments support this idea. First, because respondents in conjoint experiments are typically presented with a large number of features, the design allows respondents to justify any particular choice or rating (Hainmueller, Hopkins, and Yamamoto 2014). Secondly, due to the large number of varying features, the main research objective of the study is unclear to respondents (Hainmueller, Hopkins, and Yamamoto 2014; Ono and Yamada 2016). For example, in a study of gender biases in voters' evaluations of political candidates, Teele, Kalla, and Rosenbluth (2018) state that because candidate gender is embedded as one of multiple features *'(...) our own interest in gender would not have been obvious in the experiment. This likely lessens the degree to which our results are skewed by social desirability bias'*.

The notion that researchers can mitigate SDB and obtain more reliable answers when research intentions are 'masked' is not new. Previous survey research on sensitive topics have implemented cover stories in order to misdirect participants about the goal of the experiment (McDermott 2002; Dickson 2011). For example, by asking questions unrelated to the primary intention of the study (Kam 2007) or by providing respondents with an alternative or vaguely stated purpose of the experiment (Bullock 2011; Arceneaux 2008).

The arguments for why conjoint designs should minimize concerns over SDB appear plausible,

[5] A researcher demand effect is distinct from SDB and happens when respondents infer the response researchers expect and behave in line with these expectations (Mummolo and Peterson 2018). In principle, demand effects could work in the opposite direction of SDB which I test in the final part of the paper.

but there is little empirical evidence to support them. On the one hand, some studies that use conjoint designs have implemented various tests in order to reject that SDB is an issue. For example, Bansak, Hainmueller, and Hangartner (2016) find that results are stable for respondents with different levels of empathy, building on the idea that empathy and social desirability scales correlate. Hainmueller and Hopkins (2015) come to the same conclusion after re-estimating their results based on measures of self-monitoring that are known to be closely connected to social desirability. Finally, Hainmueller, Hangartner, and Yamamoto (2015) use a natural experiment as a behavioral benchmark and compare the results from conjoint experiments with real-world behavior.

On the other hand, results from several conjoint experiments that study sensitive dimensions seem at odds with what we know from field experiments or observational studies and run counter to observed real-world outcomes. For example, a number of studies on voter preferences that use candidate conjoint designs find no effects – or even positive effects – of being a non-white political candidate compared to a white political candidate (Carnes and Lupu 2016; Kirkland and Coppock 2017).[6] These results contradict studies of actual voting patterns (e.g. Broockman and Soltas (2017) and Lewis-Beck, Tien, and Nadeau (2010)). This seems to indicate that the results from conjoint experiments may be biased because of SDB. This concern is further strengthened by recent evidence suggesting that experimental findings on voter preferences for women or black candidates may overestimate support, even in anonymous settings (Krupnikov, Piston, and Bauer 2016).

Finally, while Hainmueller, Hangartner, and Yamamoto (2015) demonstrate that the paired-conjoint design is aligned with real-world behavior, they also demonstrate that *'seemingly subtle differences in survey designs can produce significant differences in performance'*. In summary, there is reason to suspect that SDB can be an issue in conjoint experiments, making it pertinent to understand if design adjustments can mitigate this type of response bias.

[6] Carnes and Lupu (2016) conducted a conjoint experiment in which they manipulated candidates' race using two levels (white and black) in a study of support for political candidates, and find a positive (although only borderline significant) effect of being black. Similarly, Kirkland and Coppock (2017) finds that Hispanic, Black and Asian candidates respectively are preferred over White candidates (although these differences are not significant).

Research design

I conducted two independent studies each comprising three conjoint experiments specifically designed to assess the relation between design and SDB. The experiments are almost identical to two previous studies by Sen (2017) and Hainmueller and Hopkins (2015).[7] The experiments were implemented in Qualtrics software and fielded in August 2018 on a total of 7,059 respondents recruited from Amazon's Mechanical Turk, which hosts an experienced pool of survey respondents (Berinsky, Huber, and Lenz 2012).[8]

Manipulating attention to sensitive features through design

Both studies include a feature that is known to be influenced by social desirability pressures. Study 1 seeks to gauge the effect of candidates' race, a topic to which it can be difficult to obtain honest self-reports since racial preferences is taboo (Krupnikov, Piston, and Bauer 2016; Berinsky and Lavine 2012). Study 2 seeks to explore support for immigrants seeking admission to the US. Religious affiliation, more specifically being Muslim, serve as a sensitive feature level. Restricitionist immigration policies is a hot-button topic that previous research has found to be subject to SDB (Janus 2010).

We can think of the identification strategy as a two-stage process. The first stage concerns the link between the specific design and respondents' attention to sensitive features and their possibility of justifying inappropriate answers. The second stage concerns whether this affect respondents' priorities. I seek to manipulate respondents' awareness to the sensitive feature in two ways. First, I manipulate the probability weights of the levels of the sensitive feature across conditions. Thus, one condition, the *high-contrast design*, is a paired-conjoint in which respondents are presented with five different candidate pairs with each or most pairs displaying a contrast on the sensitive feature

[7] The designs in the present study differ slightly from the original studies in terms of the number and type of features included. Considering that the purpose of the present study being not to replicate these studies, but rather to determine whether treatment effects vary across design, this is not problematic.

[8] In 2018, researchers raised concerns that an increasing number of "bots" (respondents using semi- or fully-automated code to automatically respond) reduced the quality of answers to surveys fieled on Amazon's Mechanical Turk. In order to weed out potential bots I used reCAPTCHA and a basic quality check (What is 2+2?).

(for example, a black vs. a white candidate). Arguably, the repeated contrast increase respondents' awareness to the sensitive feature. Moreover, the frequent contrast makes it harder for respondents to defend an inappropriate answer since they have to repeat it across five conjoint pairs. We would expect SDB to amplify in this condition. In the second condition, the *restricted paired design*, the sensitive feature is contrasted less frequently. Thus, the restriction serves to mask the sensitive feature from respondents by design.

Secondly, I test the importance of the within-subject structure that characterizes the paired design by including a *single-profile design* as a third condition (See details on the conjoints in appendix A). While respondents in a paired conjoint design observe *both* treatment and control at the same time, the single-profile conjoint displays *either* control or treatment which arguably makes the sensitive feature less noticeable. Again, I expect this design to reduce social desirability pressures compared to the high-contrast paired design.

In each study, respondents are randomly assigned to one of the three conditions. Because the second condition is restricted on the sensitive feature which reduces statistical power, half of the respondents are assigned to this condition in order to gain precision, while a quarter of the sample is assigned to the high-contrast design and the single-profile design respectively.

Study 1: U.S. Supreme Court nominees

The first study is inspired by a candidate conjoint study on support for Supreme Court nominees by Sen (2017). The design is a typical example of a conjoint design in which the researcher asks a sample of 1,650 U.S. adults to rank a number of hypothetical candidates. While the original study used three different outcome measures on a 7-point likert scale ("Support", "Qualifications", and "Trust"), I only ask respondents to either indicate who they support most or, in the single-profile, to rate their level of support for the candidate. Also, I exclude information on political leaning that was assigned to half of the respondents in the original study.

Respondents are randomly assigned to one of the three conjoint experiments that are otherwise identical in terms of features, levels, wording and formatting. The experiments include six features that each hold several feature levels (See details in appendix B). Most importantly, candidates' race

are assigned from a list with two levels (black or white).

In the high-contrast condition respondents are presented with five different pairs in which each pair contrasts candidates' race. That is, all five candidate pairs appear as *Black vs. White* or *White vs. Black*. The second condition is equivalent to the first except candidates' race is restricted to appear only in one of the five pairs. In the final condition, respondents are presented with a single-profile conjoint in order to eliminate the contrast on race that is inherent to the comparison in paired-conjoint designs. In this condition the candidates' race is assigned randomly. The design is summarized in Figure 1.

Study 2: Immigrants seeking admission to the U.S.

Study 2 is substantively inspired by Hainmueller and Hopkins (2015) and examines respondents' support for immigrants applying for admission to the U.S.[9] As in Study 1, respondents are assigned to one of three variations of a conjoint design that all include seven features (See details in appendix B). Most importantly here is *Religion* that can take on six levels (Catholic, Protestant, Jewish, Muslim, Atheist or Other). I follow roughly the same strategy as in study 1, and assign respondents to three different conjoint designs varying the focus on the sensitive feature. In the high-contrast design, the probability that one of the two profiles in any given pair is Muslim is high (80 per cent of all pairs), whereas in the second condition the probability that one of the two profiles is Muslim is restricted (17 per cent of all pairs). Finally, in a single-profile conjoint, religious affiliation is drawn randomly, but as was the case in study 1, the religious contrast is arguably not as prominent due to the non-paired structure of the design.

[9] In the original study, the features were chosen to approximate the information available to immigration officials which is why religion was omitted, but the authors suggest religion as a dimension for future work to explore.

Figure 1: Experimental conditions in Study 1 and Study 2

	Paired conjoint, High contrast	**Paired conjoint, Restricted**	**Single-profile conjoint**
Study 1 Feature: Race Levels: *Black / White*	5 of 5 pairs contrast race **N = 854**	1 out of 5 pairs contrasts race **N = 1765**	Random assignment of race **N = 874**
Study 2 Feature: Religion Levels: *Muslim / Protestant / Catholic / Jewish / Atheist / Other*	80 % chance that one candidate is Muslim **N = 926**	17 % chance that one candidate is Muslim **N = 1770**	Random assignment of religion **N = 870**

Results

Can respondents infer research intentions?

A first-order concern is whether the design variations in fact have an effect on respondents' awareness to the sensitive feature. To check if this is the case, the survey included a post-treatment question asking respondents to choose from a list of eight different options what they believed to be the main objective of the study. As displayed in Figure 2 respondents' anticipation of the research objective changes drastically across design conditions. In Study 1, 38 percent of respondents in the high-contrast condition answered that the primary intent of the study was to examine their reactions to candidates' race. This is 30 percentage points more compared to the restricted paired conjoint, and 13 percentage points more relative to the single-profile conjoint. We see the same pattern in Study 2. 48 percent of respondents in the high-contrast paired design believed that the main objective of the study was to examine support for immigrants conditional on their religious affiliation, which is 20 percentage points more than in the restricted paired conjoint and 16 percentage points more compared to the single-profile conjoint.

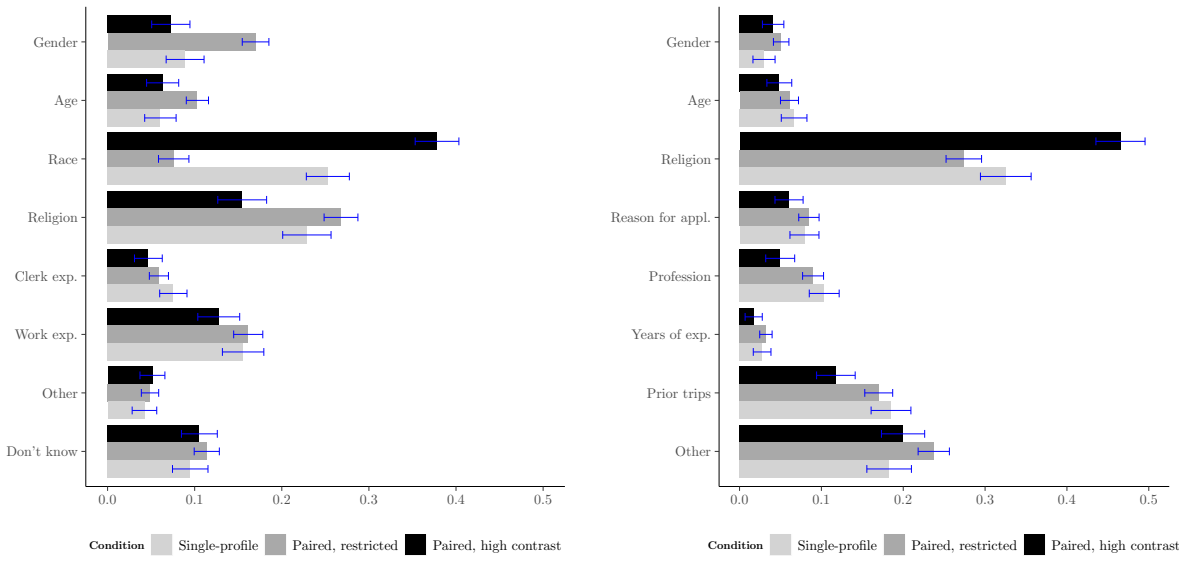
This demonstrates two important points. First, that respondents pay much attention to sensitive features such as race or religious affiliation when answering these experiments. Second, that seemingly subtle design differences significantly affect respondents' inferences about research inten-

tions. In other words, it *is* possible to downplay a sensitive dimension by adjusting the design and thus make respondents significantly less likely to infer that the sensitive feature is important.

Figure 2: Manipulation check. Respondents perception of the main research objective

(a) Study 1. US Supreme Court Nominees

(b) Study 2. Immigrants seeking admission



Note: The figures indicate the distribution of respondents' anticipation of the main research objective across the three designs. Figure (a) at the left depicts the results from study 1, while figure (b) at the right shows the results from study 2.

Building on the common assumption that respondents give more desirable answers when they anticipate that a sensitive feature is the main research objective, we would expect respondents to give different answers across conditions. More specifically, respondents should be more favorable to the black political candidates as well as the Muslim immigrants in the high-contrast design relative to the restricted paired design and the single-profile design.

Does design variation affect respondents' behavior?

Before statistically testing the differences across designs, the AMCEs from the high-contrast designs are compared with the alternative designs in four scatterplots.[10] A traditional visualization of the results from each study including attribute level-names are reported in Appendix C. Notice that the single-profile conjoint designs rely on a different type of task (evaluating one profile at a time instead of choosing between two) and a rating-based outcome measure. When analyzing the single-profile design, I use the ratings to code a binary variable as 1 if the rating is above the midpoint and 0 otherwise as is standard in the literature (Hainmueller and Hopkins 2015). This implies that the unweighted effect estimates are not immediately comparable with the paired designs. Yet, the magnitude of the AMCE of the sensitive feature levels relative to the other AMCEs is directly comparable in the scatterplots.

Figure 3 depicts the results from Study 1. The left plot shows each coefficient estimate for the high-contrast design versus estimates obtained from the restricted design. The right plot shows each coefficient estimate from the high-contrast design versus estimates obtained from the single-profile design. Each point represents an AMCE-estimate with 95 percent confidence intervals with the coefficients ordered by their magnitude from most negative to most positive. Thus, the figure visualizes the extent to which larger AMCEs in the high-contrast designs are associated with larger effects in the alternative designs. In the same way, Figure 4 compares the estimates from study 2 when comparing the high-contrast design versus each estimate obtained from the restricted design (left side) and the single-profile design (right side). Altogether, there are no apparent differences in the AMCEs between the experimental conditions. Generally, the different designs yield highly comparable results. Importantly, this is also the case when comparing the AMCE-estimates of the sensitive features across designs (estimates are colored blue in the figures).

[10] The analysis was conducted using R package version 0.3.1 (Leeper 2019).

Figure 3: Study 1. High-contrast estimates versus estimates from the alternative designs

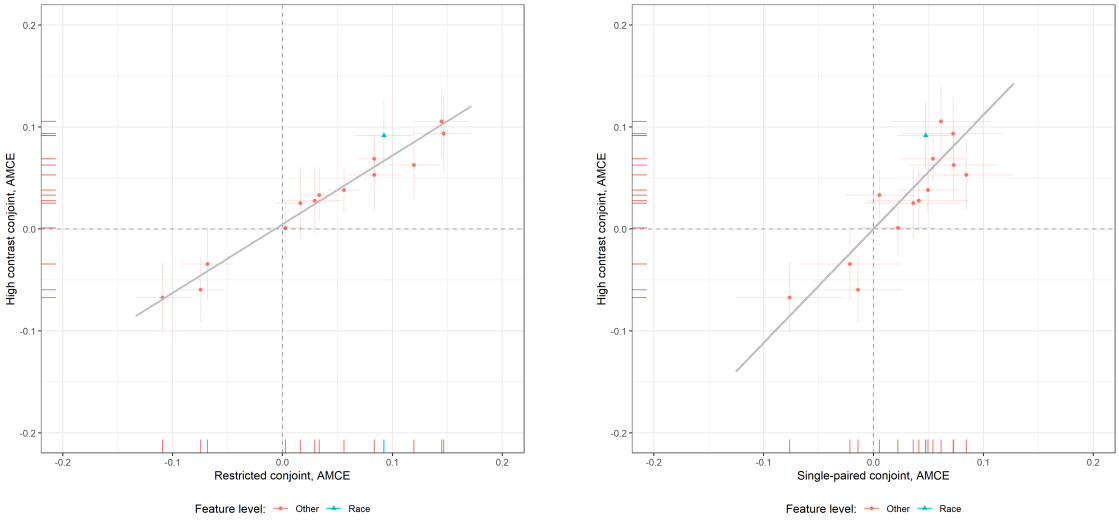
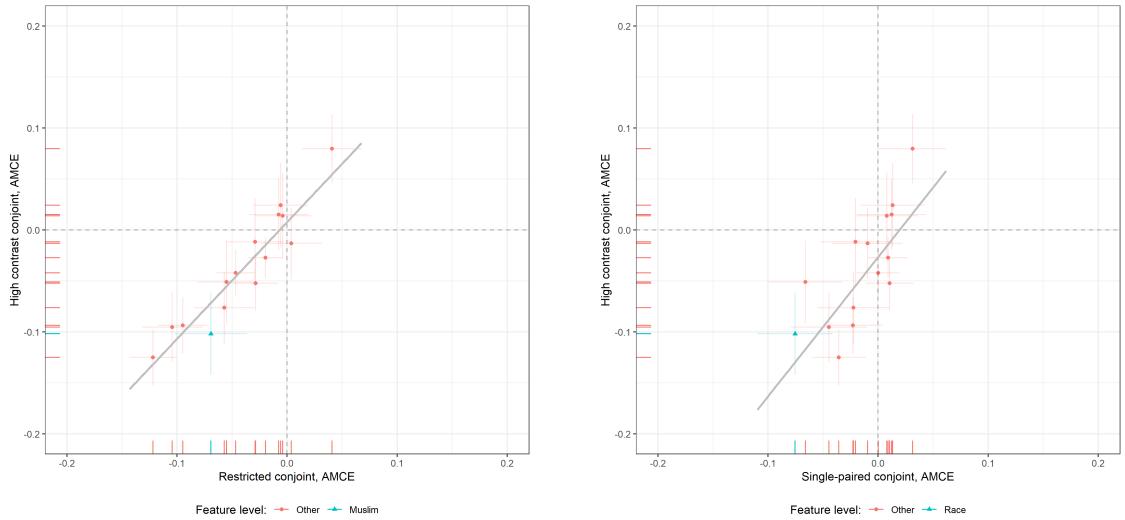


Figure 4: Study 2. High-contrast estimates versus estimates from the alternative designs



Note: The figures show each AMCE-estimate from the high-contrast design versus estimates obtained from the restricted (left) and the single-profile designs (right). Each point represents an AMCE-estimate with 95 percent confidence intervals with the coefficients ordered by their magnitude from most negative to most positive.

Next, I turn to a statistical comparison of the results. The outcome of interest is the differences in effects of the sensitive feature levels across designs. First, the two paired conjoint experiments in each study are compared. The paired designs rely on the same outcome and are therefore directly comparable. Hence, the effect of reducing attention to the the sensitive topic can be tested in a difference-in-difference model. In other words, I interact a design dummy variable (high-contrast = 0 / restricted = 1) with the sensitive topic in each study respectively.[11] A positive estimate indicates that respondents give more desirable answers in the high-contrast design which aligns with the expectation that SDB can be introduced by raising awareness to the sensitive feature. As shown in Figure 5 (a) the difference in the effects of the sensitive feature between designs is remarkably close to zero in both studies.[12] The effect of being black (Study 1) or Muslim (Study 2) is identical across designs.

In Figure 5 (b) I follow the same strategy in order to compare the high-contrast paired design and the single-profile design. However, since the experiments rely on outcomes measured on different scales, the comparison is not as straightforward. Since the AMCEs are consistently smaller in the single-profile designs, the size of the effect of a candidate being Black or Muslim is naturally smaller compared to the paired designs. I account for this by weighting the AMCEs in the single-profile design using the relative difference of all AMCEs between the single-profile and the paired designs as a weight.[13] Again, there are no substantial differences between the single-profile and the high-contrast paired designs either as evidenced from Figure 5 (b).[14]

---

[11] For example, the estimand comparing the two paired design is expressed as:

$$(E[choice \mid Black \quad \& \quad High\text{-}contrast] - E[choice \mid White \quad \& \quad High\text{-}contrast) \\ -(E[choice \mid Black \quad \& \quad Restricted] - E[choice \mid White \quad \& \quad Restricted])$$

An equivalent estimand is used in study 2 where "Black" equals "Muslim" and "White" equals the reference category.

[12] In study 1, the effect of being black compared to white increases the probability that a profile is chosen by 0.089 (SE = 0.016) in the high-contrast design and 0.092 (SE = 0.013) in the restricted paired conjoint. The effect of being Muslim is negative in both the high-contrast design with a coefficient of -0.096 (SE = 0.020) and -0.070 (SE = 0.017) in the restricted paired design.

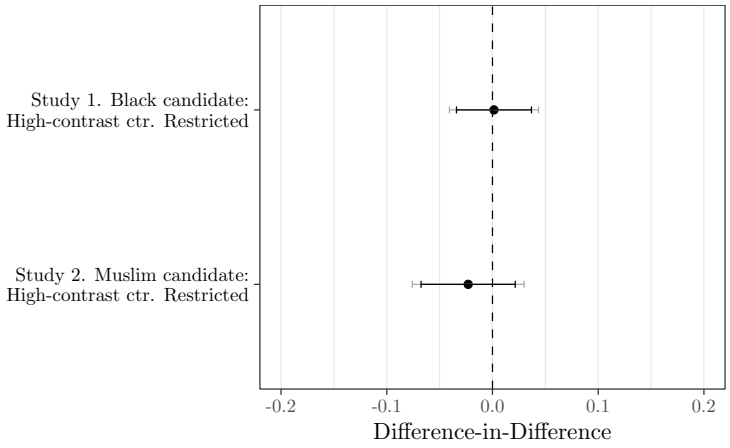
[13] Specifically, the paired designs give AMCEs that are on average larger by a factor 1.87.

[14] Note that the difference-in-differences are insignificant also without weighting the AMCEs from the Single-profile. See details in appendix F.

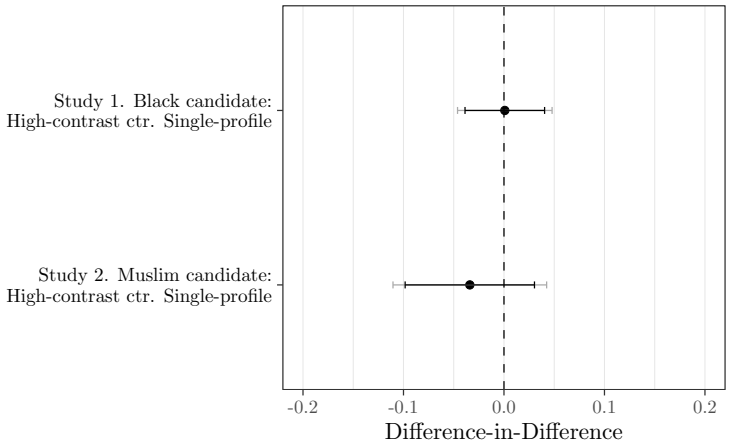
In summary, the results show that even when respondents anticipate a sensitive feature as important and at the same time have optimal conditions for tailoring their answers, it does not change their responses.

Figure 5: Difference-in-differences

(a) High-contrast paired design vs. Restricted paired design



(b) High-contrast paired design vs. Single-profile design



Note: The figures show differences in effects of the sensitive levels between (a) the high-profile paired design and the restricted paired design and (b) the high-profile paired design and the single-profile design.

What else could explain the null-findings?

The results are supposedly good news to researchers conducting conjoint experiments: we should not be too concerned with implementing designs that, at least in theory, increase the risk of SDB. In this section I test alternative explanations for the null-findings.

First and foremost, one concern is that (some) respondents would disagree that it is socially desirable to have a preference against black or Muslim candidates.[15] Another methodological objection is that the increased awareness to a sensitive feature also introduces demand-effects that cancels out SDB. Demand effects are caused by respondents attempting to validate a researcher's hypothesis by behaving in line with what they perceive as the expected behavior (Mummolo and Peterson 2018). If respondents anticipated that the present study expected to find a bias against black or Muslim profiles, they might have answered in a way that would "help" the researcher confirm the hypothesis, which would bias the effect in the opposite direction than SDB.

To bolster the results, I therefore rerun the analysis in subsets of the samples where social desirability pressures related to preferences regarding Race and Muslim affiliation are arguably more pronounced. Firstly, political liberals have been found to be more likely to give untruthful answers to questions regarding race (Gilens, Sniderman, and Kuklinski 1998) and immigration restrictionist policy questions (Janus 2010). In both studies, I reestimate the difference-in-differences between the paired designs in subsets of respondents that identify as liberal on the pre-treatment questions.[16] As demonstrated in the supplementary material (appendix D), the difference-in-differences from the liberal subset is a precisely estimated zero (Study 1: -0.014, SE = 0.031; Study 2: 0.0007, SE = 0.049). Secondly, to further bolster the results, I look at subsets of the samples that are more likely to be attentive to self-presentational concerns – and thus where we would expect SDB to be most pronounced. Previous studies have found that high self-monitors are more likely to give appropriate answers to sensitive questions. Following Berinsky and Lavine (2011), study 2 included three items from the self-monitoring scale that was also used in a conjoint analysis by Hainmueller and Hopkins

[15] Although M-turkers tend to be younger and more liberal compared to a nationally representative samples (Levay, Freese, and Druckman 2016).

[16] Respondents with a score >6 on a 0-10 scale ranging from "Very conservative" to "Very liberal".

(2015).[17] As shown in the supplementary material (appendix D), the difference-in-differences from the high-monitor subset is close to null in study 2 (.020, SE = 0.043).

A third concern is, that the "treatment" in the paired design with high contrast was not assigned before the outcomes were measured, but rather is embedded in the design. Hence, it is possible that the sensitive dimension became increasingly obvious to respondents as they worked their way through the five conjoint pairs. In other words, respondents assigned to the high-contrast paired design could have been more aware about the sensitive feature when they where asked to choose between a black and a white candidate for the third, fourth and fifth time. In that case, results should change towards more politically correct answers towards the end of the experiment. To test this, I compare estimates in the high-contrast designs from pair 1-5 respectively. The change in effect sizes as respondents answer the five pairs in the high-contrast designs are inconsequential and does not support the notion that respondents change preferences as the contrast on a sensitive dimension is repeated (See appendix D in the supplementary material).

Finally, we might worry that the semi-professional respondents on Amazon's Mechanical Turk are somehow less prone to social desirability pressures than population based samples. The experiments provided in this study cannot shed light on this concern. However, research indicates that survey experiments conducted on convenience samples like M-turkers yield similar effects as those from national probability samples (Berinsky, Huber, and Lenz 2012; Coppock 2018). Furthermore, since a lot of social science conjoint experiments are carried out in convenience respondent pools such as M-turk, examining the research question in a convenience sample has a value in itself.

[17] The following questions are used: "When you're with other people, how often do you put on a show to impress or entertain them?" Response categories: Always, Most of the time, About half the time, Once in a while, Never. "How good or bad of an actor would you be?" Response categories: 'Excellent', 'Good', 'Fair', 'Poor', 'Very poor'. "When you are in a group of people, how often are you the center of attention?" Response categories: 'Always', 'Most of the time', 'About half the time', 'Once in a while', 'Never'.

Conclusion and discussion

Conjoint designs are often claimed to limit concerns over social desirability bias: that research subjects are biased towards normatively positive responses. This is based on two arguments: due to the large number of features, (a) respondents cannot infer the main intent of the experiment, and (b) they can easily justify inappropriate answers. However, the extent to which these arguments hold depends on the specific type of conjoint experiment employed. The present study tests the importance of design variations by comparing answers across different types of conjoint designs.

The results provide evidence that the design of conjoint experiments have an effect on respondents' inferences about the main objective: respondents pay significantly more attention to sensitive features in a paired conjoint design where the sensitive feature levels are frequently contrasted compared to designs where the contrast is less obvious (single-profile and paired conjoints with restricted randomization schemes). However, the core quantities of interest are remarkably stable across designs, suggesting that the substantive conclusions are not threatened by the specific choice of design.

There are several implications of these results. First, while this study cannot rule out that SDB is ever an issue in conjoint experiments, it is reassuring that different types of conjoint designs give the same results. Second, the stability of the results across designs also goes against recent suggestions that paired-conjoint designs makes it easier for respondents to act 'strategically' in order to provide desirable answers (Mummolo and Peterson 2018). There is no evidence that respondents act differently when presented with a within-subject design compared to a between-subject study. Thirdly, and consequently, there is no immediate reason to choose a design that is otherwise sub-optimal in order to disguise sensitive topics. Had this study proved that respondents' priorities change when respondents recognize sensitive features as important, the implications would be serious. As a main concern, it would question the inferences that researchers are able to make from conjoint designs more generally. Moreover, it would emphasize the need to choose otherwise sub-optimal designs in order to downplay sensitive features. The results presented in this article indicates that there is no reason that researchers using conjoint designs should limit the number

of pairs or restrict the probabilities of certain feature levels and thereby decrease statistical power and/or hamper external validity. Nor is there strong arguments for using single-profile designs unless they are preferable for other reasons. Finally, beyond conjoint designs specifically, the present study supports recent evidence by Mummolo and Peterson (2018) and De Quidt, Haushofer, and Roth (2017) that researchers should not be too concerned with respondents' awareness to research intentions in survey experiments.

References

Arceneaux, Kevin (2008). "Can partisan cues diminish democratic accountability?" In: *Political Behavior* 30.2, pp. 139–160.

Bansak, Kirk, Jens Hainmueller, and Dominik Hangartner (2016). "How economic, humanitarian, and religious concerns shape European attitudes toward asylum seekers". In: *Science*, aag2147.

Bansak, Kirk, Jens Hainmueller, Daniel J Hopkins, et al. (2017). "Beyond the breaking point? Survey satisficing in conjoint experiments". In: *Political Science Research and Methods*, pp. 1–19.

Berinsky, Adam J, Gregory A Huber, and Gabriel S Lenz (2012). "Evaluating online labor markets for experimental research: Amazon. com's Mechanical Turk". In: *Political analysis* 20.3, pp. 351–368.

Berinsky, Adam J and Howard Lavine (2012). "Self-monitoring and political attitudes". In: *Improving public opinion surveys: Interdisciplinary innovation and the American national election studies*, pp. 27–45.

Broockman, David and Evan Soltas (2017). *A natural experiment on taste-based racial and ethnic discrimination in elections*.

Bullock, John G (2011). "Elite influence on public opinion in an informed electorate". In: *American Political Science Review* 105.3, pp. 496–515.

Carnes, Nicholas and Noam Lupu (2016). "Do voters dislike working-class candidates? Voter biases and the descriptive underrepresentation of the working class". In: *American Political Science Review* 110.4, pp. 832–844.

Coppock, Alexander (2018). "Generalizing from survey experiments conducted on Mechanical Turk: A replication approach". In: *Political Science Research and Methods*, pp. 1–16.

De Quidt, Jonathan, Johannes Haushofer, and Christopher Roth (2017). *Measuring and Bounding Experimenter Demand*. Tech. rep. National Bureau of Economic Research.

Dickson, Eric (2011). "Economics vs. Psychology Experiments". In: *The Handbook of Experimental Political Science*. Cambridge University Press.

Gilens, Martin, Paul M Sniderman, and James H Kuklinski (1998). "Affirmative action and the politics of realignment". In: *British Journal of Political Science* 28.1, pp. 159–183.

Groves, Robert M et al. (2011). *Survey methodology*. Vol. 561. John Wiley & Sons.

Hainmueller, Jens, Dominik Hangartner, and Teppei Yamamoto (2015). "Validating vignette and conjoint survey experiments against real-world behavior". In: *Proceedings of the National Academy of Sciences* 112.8, pp. 2395–2400.

Hainmueller, Jens and Daniel J Hopkins (2015). "The hidden American immigration consensus: A conjoint analysis of attitudes toward immigrants". In: *American Journal of Political Science* 59.3, pp. 529–548.

Hainmueller, Jens, Daniel J Hopkins, and Teppei Yamamoto (2014). "Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments". In: *Political Analysis* 22.1, pp. 1–30.

Hariri, Jacob Gerner and David Dreyer Lassen (2017). "Income and outcomesSocial desirability bias distorts measurements of the relationship between income and political behavior". In: *Public Opinion Quarterly* 81.2, pp. 564–576.

Holtgraves, Thomas (2004). "Social desirability and self-reports: Testing models of socially desirable responding". In: *Personality and Social Psychology Bulletin* 30.2, pp. 161–172.

Horiuchi, Yusaku, Daniel M Smith, and Teppei Yamamoto (2017). "Identifying Voter Preferences for Politicians' Personal Attributes: A Conjoint Experiment in Japan". In:

Janus, Alexander L (2010). "The influence of social desirability pressures on expressed immigration attitudes". In: *Social Science Quarterly* 91.4, pp. 928–946.

Kam, Cindy D (2007). "Implicit attitudes, explicit choices: When subliminal priming predicts candidate preference". In: *Political Behavior* 29.3, pp. 343–367.

Kaminska, Olena and Tom Foulsham (2013). *Understanding sources of social desirability bias in different modes: Evidence from eye-tracking*. Tech. rep. ISER Working Paper Series.

Kirkland, Patricia A and Alexander Coppock (2017). "Candidate Choice Without Party Labels". In: *Political Behavior*, pp. 1–21.

Krupnikov, Yanna and Adam Seth Levine (2014). "Cross-sample comparisons and external validity". In: *Journal of Experimental Political Science* 1.1, pp. 59–80.

Krupnikov, Yanna, Spencer Piston, and Nichole M Bauer (2016). "Saving Face: Identifying Voter Responses to Black Candidates and Female Candidates". In: *Political Psychology* 37.2, pp. 253–273.

Krysan, Maria and Mick P Couper (2003). "Race in the live and the virtual interview: Racial deference, social desirability, and activation effects in attitude surveys". In: *Social psychology quarterly*, pp. 364–383.

Kuklinski, James H et al. (1997). "Racial prejudice and attitudes toward affirmative action". In: *American Journal of Political Science*, pp. 402–419.

Leeper, Thomas J. (2019). *cregg: Simple Conjoint Analyses and Visualization*. R package version 0.3.1.

Levay, Kevin E, Jeremy Freese, and James N Druckman (2016). "The demographic and political composition of Mechanical Turk samples". In: *Sage Open* 6.1, p. 2158244016636433.

Lewis-Beck, Michael S, Charles Tien, and Richard Nadeau (2010). "Obama's missed landslide: a racial cost?" In: *PS: Political Science & Politics* 43.1, pp. 69–76.

Liu, Hanzhang (2018). "The Logic of Authoritarian Political Selection: Evidence from a Conjoint Experiment in China". In: *Political Science Research and Methods*, pp. 1–18.

McDermott, Rose (2002). "Experimental methods in political science". In: *Annual Review of Political Science* 5.1, pp. 31–61.

Mummolo, Jonathan and Erik Peterson (2018). "Demand effects in survey experiments: An empirical assessment". In: *American Political Science Review*, pp. 1–13.

Ono, Yoshikuni and Masahiro Yamada (2016). "Do Voters Prefer Gender Stereotypic Candidates?: Evidence from a Conjoint Survey Experiment in Japan". In:

Sen, Maya (2017). "How political signals affect public support for judicial nominations: Evidence from a conjoint experiment". In: *Political Research Quarterly* 70.2, pp. 374–393.

Teele, Dawn Langan, Joshua Kalla, and Frances Rosenbluth (2018). "The Ties that Double Bind: Social Roles and Women's Underrepresentation in Politics". In: *American Political Science Review*, pp. 1–17.

Tourangeau, Roger and Ting Yan (2007). "Sensitive questions in surveys." In: *Psychological bulletin* 133.5, p. 859.

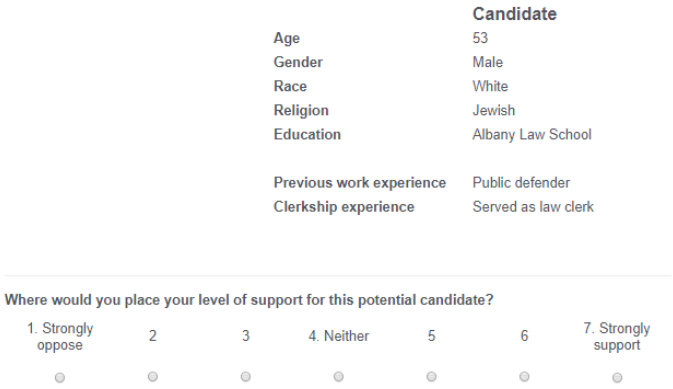
Supplementary material: Social desirability bias in conjoint experiments: What is the optimal design when studying sensitive topics?

Appendix A. Constructing and fielding the conjoint experiments

The experiments were implemented in Qualtrics and fielded at Amazon's Mechanical Turk. The sampling took place between August 8 and August 30. The sampling design was a random sampling using the build-in randomize option in Qualtrics. Only respondents who answered the last question (the manipulation check) are included in the final sample. The respondents were presented with a paired design or a single-profile design. Screenshots of a paired conjoint design and a single-profile conjoint design are shown in Figure A1 and A2.

A1. Example of discrete choice conjoint

A2. Example of rating-based conjoint



Appendix B. Features and feature levels

Table B give details on the features and feature levels used to generate the profiles in Study 1 and Study 2.

Study 1		Study 2	
Feature	Level	Feature	Level
Gender	Male, Female	Gender	Male, Female
Age	25-75 (Continuous)	Age	25-75 (Continuous)
Race	White, Black	Religion	Atheist, Protestant, Jewish, Muslim, Catholic, Other
Education	Yale Law School, Florida State Uni., Albany Law School	Reason for application	Seek better job, Reunite with family members, Escape religious/political persecution
Religion	Mormon, Mainline Protestant, Jewish, Evangelical Protestant, Catholic	Profession	Doctor, Nurse, Teacher, Waiter, Construction worker, Computer programmer
Clerkship experience	Did not serve as law clerk, did serve as law clerk	Working experience	None, 1-2 years, 3-5 years, More than 5 years
Previous work experience	elected politician, law professor, lawyer in private practice, non-profit lawyer, public defender	Prior trips to the U.S.	Never been to the U.S., Spent six months with family, Visited once without legal authorization, Visited once on tourist visa, Visited many times on tourist visa

B. Features and feature levels included in the conjoint experiments

Appendix C. Results from conjoint experiments

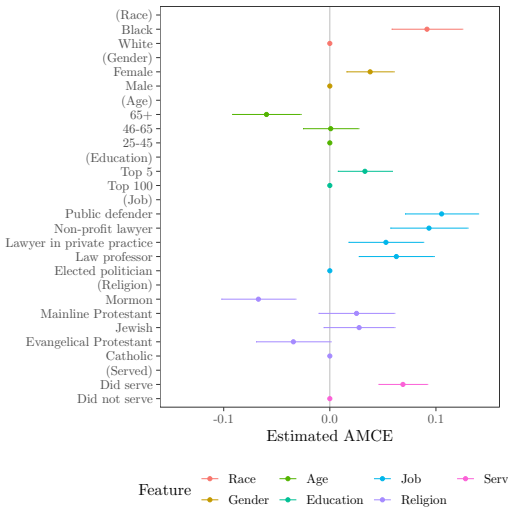
The AMCEs from each study are visualized in Figure 6.[18] Notice that the single-profile conjoint designs rely on a different type of task (evaluating one profile at a time instead of choosing between two) and a rating-based outcome measure. When analyzing the single-profile design, I use the ratings to code a binary variable as 1 if the rating is above the midpoint and 0 otherwise as is standard in the literature (Hainmueller and Hopkins 2015). This implies that the unweighted effect estimates are not immediately comparable with the paired designs. Yet, the magnitude of the AMCE of the sensitive feature levels relative to the other AMCEs are strikingly similar across the three conditions in both studies. In study 1, the effect of being black is positive and significant in all of the three conditions.[19] In study 2, the effect of being Muslim is negative and significant in all conditions. Altogether, there are no apparent differences in the core quantities of interest between the three experimental conditions.

[18] The AMCE represents the marginal effect of a given attribute averaged over the joint distribution of the remaining attributes. Standard errors are corrected for within respondent clustering.

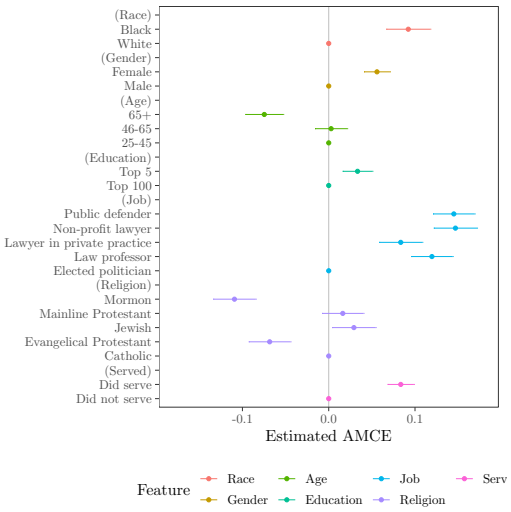
[19] This is unsurprising considering evidence from previous candidate choice experiments and the fact that this experiment did not include political leaning (which can crowd out effects of demographic characteristics).

Figure 6: Results from Supreme Court candidate conjoint experiments (N=3,493)

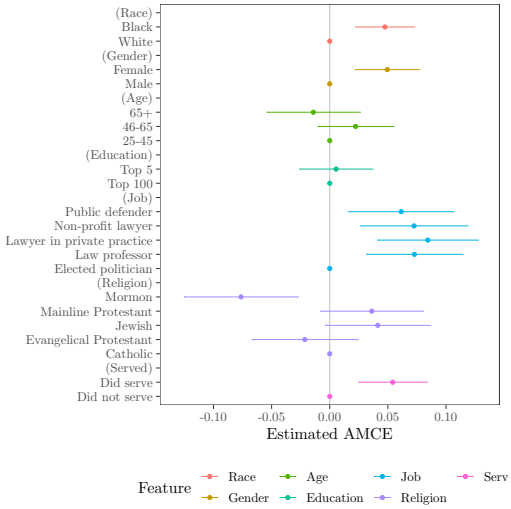
(a) High-contrast conjoint



(b) Restricted conjoint



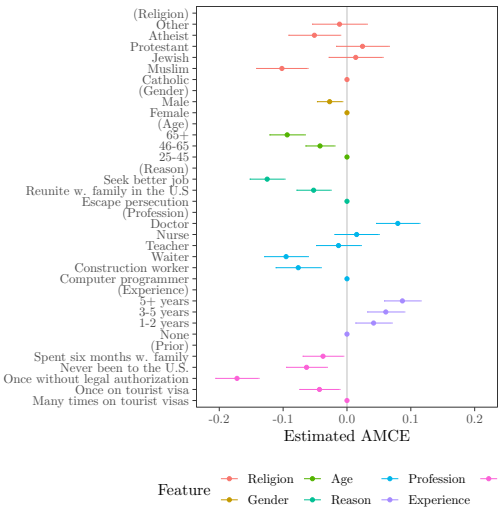
(c) Single-profile conjoint



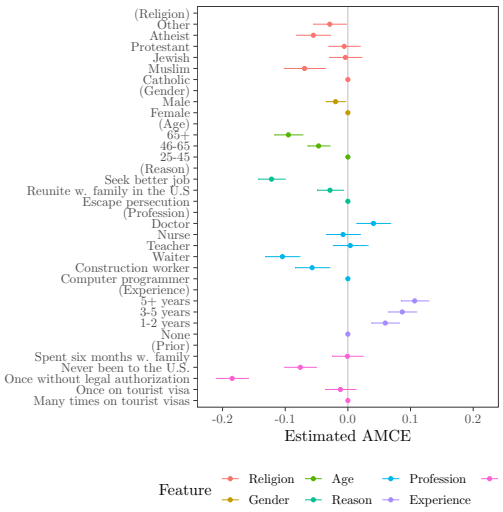
Note: Each estimate represents the effect of a given feature level compared to a reference level when averaging over the joint distribution of the remaining features.

Figure 7: Results from immigrant conjoint experiments (N=3,566)

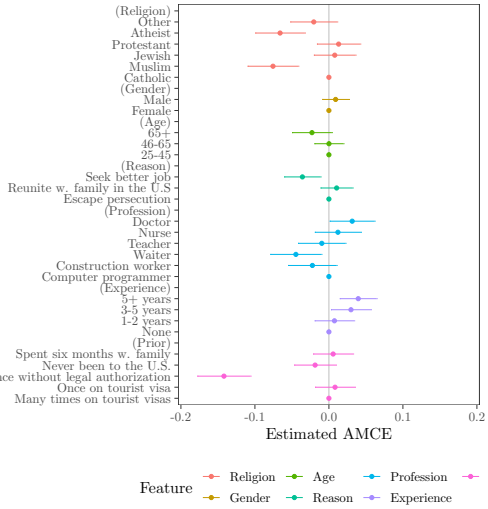
(a) Immigrant High-contrast conjoint



(b) Immigrant restricted conjoint



(c) Immigrant single-profile conjoint



Note: Each estimate represents the effect of a given feature level compared to a reference level when averaging over the joint distribution of the remaining features.

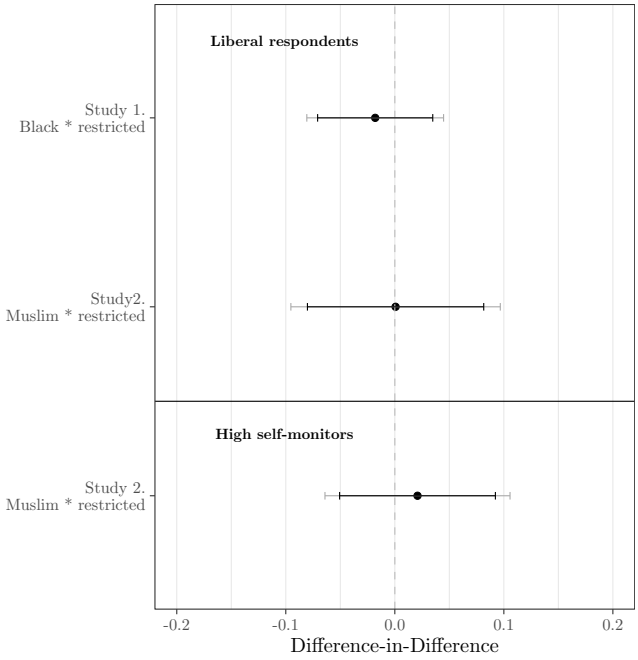
Appendix D. Robust to demand effects

One way to further bolster the results is to look at a subset of the sample that are more attentive to self-presentational concerns. First, I reestimate the difference-in-differences between the two paired designs in both studies comparing respondents across score on a 0-10 scale ranging from "Very conservative" to "Very liberal". Political liberals have been found to be more likely to give untruthful answers to questions regarding race (Gilens, Sniderman, and Kuklinski 1998) and immigration restrictionist policy questions (Janus 2010), and we would therefore expect to see stronger SDB among liberals. Secondly, previous studies have found that high self-monitors are more likely to give appropriate answers to sensitive questions. Following Berinsky and Lavine (2011), I used three items from the self-monitoring scale that was also used by Hainmueller and Hopkins (2015).[20]

[20] The following questions are used: "When you're with other people, how often do you put on a show to impress or entertain them?" Response categories: Always, Most of the time, About half the time, Once in a while, Never. "How good or bad of an actor would you be?" Response categories: Excellent, Good, Fair, Poor, Very poor. "When you are in a group of people, how often are you the center of attention?" Response categories: Always, Most of the time, About half the time, Once in a while, Never.

D. Difference-in-differences among subsets of respondents

(d) Subset of liberals and high self-monitors



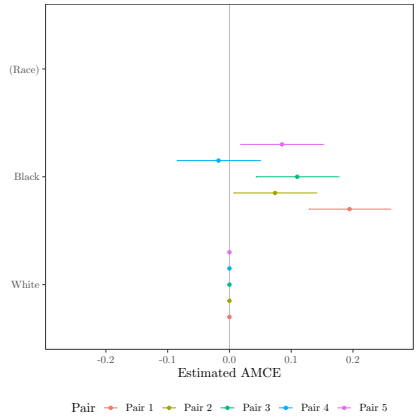
*Note:*Difference-in-differences between the paired designs when including only liberal respondents (Study 1 and Study 2) and when including only high self-monitors.

Appendix E. AMCEs across repeated choices

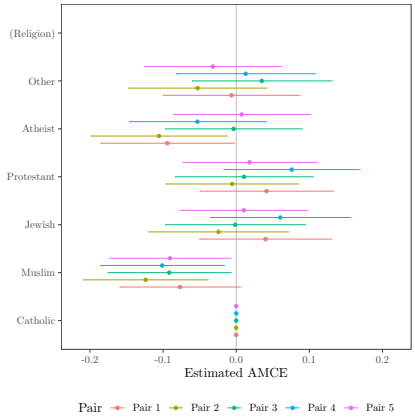
The treatment was not assigned before the experiments, but rather is embedded in the design. Hence, respondents in the paired design with high contrast may have been more aware about the sensitive feature when they where asked to choose between a black and a white candidate for the third, fourth and fifth time. To test this, I compare estimates in the high-contrast designs from pair 1, 2, 3, 4 and 5 respectively. As shown in Figure C, the change in effect sizes as respondents answer the five pairs in the high-contrast designs are inconsequential and the results do not support the notion that respondents change preferences as the contrast on a sensitive dimension is repeated.

E. Effect estimates across the five conjoint pairs

(e) Study 1. High-contrast designs conditioned on conjoint pairs



(f) Study 2. High-contrast design conditioned on conjoint pairs



Appendix F. Difference-in-difference with and without weighting

The single-profile design and the paired designs are not directly comparable. Firstly, the tasks that respondents were asked to solve differ: in the paired designs they are presented with two profiles while in the single-profile, they only see one at a time. Moreover, the outcome measure is different as well (either a forced choice or rating). Secondly, the AMCEs in the single-profile conjoint designs are generally smaller compared to the paired designs. This makes the direct comparison of the effect

estimate of the sensitive feature across design problematic. One way to solve this is to re-weight the AMCEs of the single profile designs. In other words, I estimate the average difference in effect estimates of all *other* features between the single-profile and the paired designs. On that basis, the estimate of the sensitive feature is re-weighted. Across both studies, all other features than than the sensitive are on average larger by a factor 1.87 relative to the effects in the single-profile conjoint. The paper gives the weighted difference-in-difference between the paired design with high contrast and the single-profile design. Table E gives the difference-in-difference both with and without weighting.

E1. Comparison of the high contrast paired design and the single profile designs with and without weighting

Study 2. AMCE of a candidate being black	
	Single-profile is weighted / unweighted
Paired, high contrast	0.089 (0.017)
Single-profile	0.088 (0.024) / 0.047 (0.012)
Difference-in difference	0.0008 (0.029) / 0.041 (0.022)

E2. Comparison of the high contrast paired design and the single profile designs with and without weighting

Study 2. AMCE of a candidate being Muslim	
	Single-profile is weighted / unweighted
Paired, high contrast	-0.096 (0.020)
Single-profile	-0.131 (0.033) / -0.070 (0.017)
Difference-in difference	-0.034 (0.039) / 0.026 (0.027)